Cascaded one-vs-rest detection network for fine-grained recognition without part annotations

Long Chen, *Shengke Wang, Kin-Man LAM, Huiyu Zhou, Muwei Jian, Junyu Dong

Abstract Fine-grained recognition is a challenging task due to small intra-category variances. Most of the top-performing fine-grained recognition methods leverage parts of objects for better performance. Therefore, part annotations which are extremely computationally expensive are required. In this paper, we propose a novel cascaded deep CNN detection framework for fine-grained recognition which is trained to detect a whole object without considering parts. Nevertheless, most of the current top-performing detection networks use N+1 class (N object categories plus background) softmax loss. The background category with much more training samples dominates the feature learning progress where the features are not suitable for object categorisation with fewer samples. To address this issue, we here introduce two strategies: 1) We leverage a cascaded structure to eliminate the background. 2) We introduce a novel one-vs-rest loss function to capture more minute variances from different subordinate categories. Experiments show that our proposed recognition framework achieves comparable performance against the state-of-the-art, part-free, fine-grained recognition methods on the CUB-200-2011 Bird dataset. Meanwhile, our method outperforms most of the existing part annotation based methods and does not need part annotations at the training stage whilst being free from any annotations at the test stage.

Keywords Fine-grained Recognition·Detection·One-vs-rest·Without part annotations

*Shengke Wang
Ocean University of China, Qingdao, China
e-mail: neverme@ouc.edu.cn

1 Introduction

Recently, a large body of computer vision research has focused on the fine-grained image recognition problem in several domains, such as animal breeds or species[1-3], plant species[4, 5] and architectural styles[6]. Fine-grained recognition concerns the task of distinguishing subordinate categories of the same superordinate category. It is a challenging task, as fine-grained subordinate categories share a high degree of visual similarity with small intra-class variances caused by factors such as poses, viewpoints or lighting conditions[7, 8]. Moreover, fine-grained recognition algorithms perform well within specific fine-grained domains that can provide valuable insight into a variety of challenging applications[9-13], such as the recommendation of relevant products in e-commerce, surveillance systems and so on.

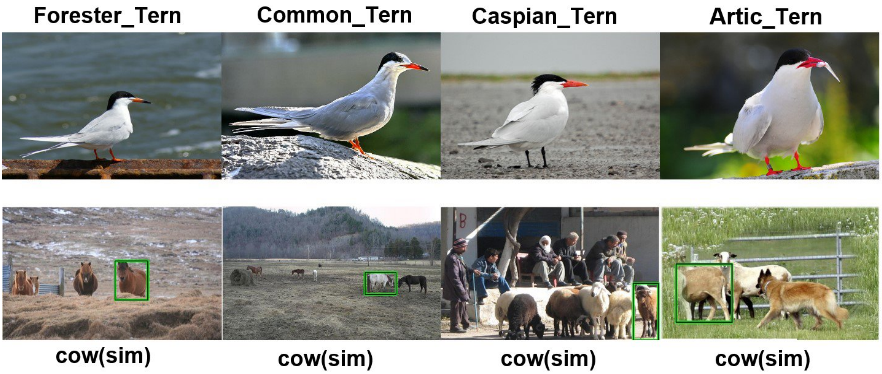

Fig.1. The top row images show the minute intra-category variances among different subordinate categories of the bird. The bottom row images show that Faster RCNN with softmax loss frequently misclassifies horses and sheep into cows, since it focus on capturing more inter-category variances rather than intra-category variances.

Most of the current state-of-the-art fine-grained recognition systems [14, 15] are part-based methods, as leveraging parts can capture the subtle appearance difference in specific object parts and achieve better performance. However, part annotations are more difficult to be obtained than object annotations. In this paper, we formulate the fine-grained recognition problem as the object detection problem[16, 17] without considering parts. When we train a standard Faster RCNN, the existence of many background samples makes the feature representation less discriminative between different subordinate categories and more confusing between an object category and the background. To address this concern, we introduce a cascaded structure to eliminate excessive background samples. Our cascaded framework consists of a standard Faster RCNN and a modified Fast RCNN with a one-vs-rest loss function. For simplicity, we denote the first standard Faster RCNN as SFNet and the unified recognition framework as RFNet. An overview of our proposed recognition framework for fine-grained recognition is shown in Fig.2. In our unified recognition framework, the standard Faster RCNN first generates primitive detections which usually contain many background parts. So we first eliminate primitive detections with low scores, which are more likely to be part of the background, and then use the balanced data to further train a modified Fast RCNN. Finally, the predicted label of the detected box with the highest score is used as the predicted label of the whole image. Our unified framework is trained to detect only the whole object, so it does not need part annotations at the training stage and is free from any annotations at the testing stage.

Fine-grained recognition tasks require distinguishing objects at the subordinate level. A good fine-grained recognition framework should be able to capture variances among different subordinate categories. However, Fast RCNN and Faster RCNN exploit the N+1 class (N object categories plus background) softmax loss function that results in an offset between detections and fine-grained recognition solutions, when referring to feature learning. The feature learning of the softmax detection network is still affected by the background class even though we have eliminated most of the background samples using the cascaded structure. Besides, it is very difficult for the softmax detection network to distinguish the objects with similar appearance or belonging to semantically related genres. For example, Faster RCNN can distinguish animals from the background, but it frequently misclassifies horses and sheep into cows (shown in Fig.1), since horse, sheep and cow are all subordinate categories of the animals and have significant intra-category variances. To bridge this gap, we replace the softmax loss function of Fast RCNN with a novel one-vs-rest loss function, which consists of N (the number of subordinate categories) two-class cross entropy losses, each of which is responsible for capturing the variances between one specific subordinate category and its similar categories. This design enables the one-vs-rest loss function to focus on capturing the variances between each category and its similar categories, suitable for fine-grained recognition tasks.

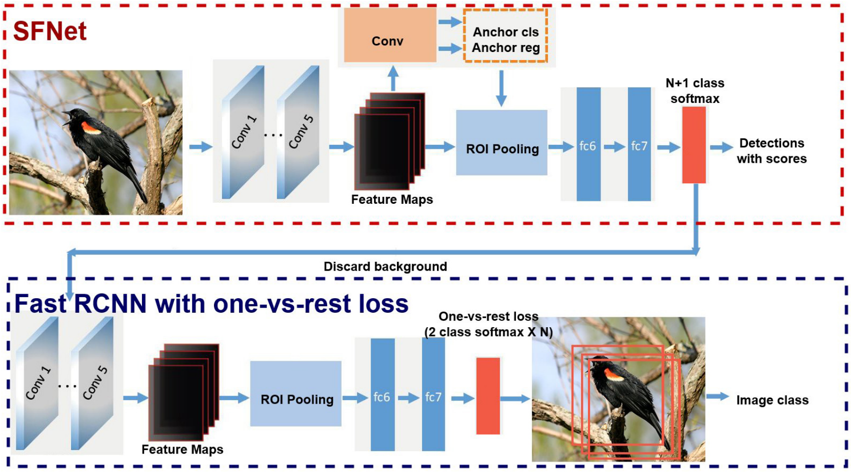

Fig.2. An overview of our RFNet. Red rectangle indicates SFNet (a standard Faster RCNN) and blue rectangle indicates one-vs-rest Fast RCNN.

The main contributions of this paper are as follows:

1) First, we propose a novel cascaded detection framework for fine-grained recognition tasks. The unified recognition framework does not need expensive part annotations at the training stage and is free from any annotations at the testing stage.

2) Second, we introduce a cascaded structure to eliminate excessive background samples, then train a better detector using the balance data. The cascaded structure enables our framework to be free from the influence of excessive background samples and the learned features are suitable for object categorisation.

3) To the best of our knowledge, it is the first time to introduce one-vs-rest detection network into fine-grained recognition tasks. Due to the ability of the one-vs-rest loss function to capture intra-category variances, the cascaded detection network is well adapted to fine-grained recognition tasks.

2 Related Work

Fine-grained recognition. Current top-performing fine-grained recognition methods [14, 15] leverage object parts, as it is widely acknowledged that the subtle difference between objects can help deliver better performance. [18, 19] focus on localizing and describing discriminative object parts in the fine-grained domain and explicitly requires both box and part annotations during the training and testing phases. Aiming at training fine-grained classifiers without part annotations, [20] introduces co-segmentation to localize the whole object and then performs alignment across all the images. [21] also leveraged better segmentation [22, 23] to localize object parts, and proposes an efficient architecture for inference, but it requires both bounding box and part annotations in training, and even needs specific annotations during testing. Towards the goal of performing fine-grained recognition without any annotations, some unsupervised methods have emerged. [24] presented a visual attention model to support fine-grained classification without any annotations. [25] reported a method to localize parts with a constellation model, which incorporates CNN into the deformable part model. Although unsupervised methods [24, 25] are free from box and part annotations, their performance is still not comparable to part-based methods. The comparison of part-based methods, bounding box-based methods and unsupervised methods can be seen in Table 1. In order to well balance the relationship between accuracy and annotation demands, we here propose a novel cascade detection framework for fine-grained recognition.

Table 1. The comparison of part-based methods, bounding box-based methods and unsupervised methods.

Methods	Advantage	Disadvantage
Part-based methods	High accuracy	Need part annotations
Box-based methods	Only need box annotations	Not accurate enough
Unsupervised methods	Without any annotations	Low accuracy

Object detection. RCNN[26] is one of the most notable region based frameworks for object detection. It demonstrates state-of-the-art performance on standard detection benchmarks at the early time and also inspires most of the state-of-the-art detection methods. RCNN first exploits the standard selective search algorithm[27] to generate hundreds or thousands of region proposals per image, and then trains a CNN to classify these region proposals. To further boost the detection performance, the standard Fast RCNN[28] and Faster RCNN[29] introduced a multi-task loss function simultaneously to classify region proposals and regress the bounding box coordinates. However, most of the current detection networks use the softmax loss function and produce a large number of misclassification errors. Recently, [30] introduced a one-vs-rest loss function in order to reduce misclassification errors in generic object detection. We here also use the one-vs-rest loss function for fine-grained recognition. Different from [30], we propose a novel cascaded detection framework for fine-grained recognition tasks and improve system performance.

3 The proposed Method

Our proposed framework consists of a standard Faster RCNN [29], followed by a modified Fast RCNN with the one-vs-rest loss function. The standard Faster RCNN first generates primitive detections which usually contain a large number of background parts. We first eliminate excessive backgrounds in the primitive detections, and then use the balanced data to further train a one-vs-rest Fast RCNN. Finally, the predicted label of the highest scored detection box is used as the predicted label of the whole image. The cascaded structure enables the one-vs-rest Fast RCNN to be free from the influence of excessive background components and the learned features are suitable for object categorisation. Besides, the softmax loss function of the Fast RCNN is replaced by a novel one-vs-rest loss function which can capture the variances between different subordinate categories.

3.1 Cascaded detection network

In order to perform fine-grained recognition without part annotations, we propose a cascaded detection framework to detect the whole object in the image so that it needs only box annotations at the training stage and is free from any annotations at the testing stage. Our cascaded framework consists of a standard Faster RCNN, followed by a one-vs-rest Fast RCNN. When training the standard Faster RCNN, the existence of many background samples allows the feature representation component to capture less intra-category variance (i.e., variance between different subcategories) and more inter-category variance (i.e., between the object category and background), causing many false positives between the ambiguous object categories (e.g., people mistakenly classify horses and sheep as cows). When training a better detector, it is necessary to eliminate excessive background samples to achieve good balance. So after eliminating the background in the primitive detections of the standard Faster RCNN, we add another one-vs-rest Fast RCNN and train it with the balanced data. The cascaded structure prevents our framework from the influence of excessive background clutters. Ref. [15] shows a Fast RCNN network to refine small semantic part candidates generated from a novel top-down proposal method, a classification sub-network to extract features from the detected parts, and combines them for recognition. In the same way, our cascaded detection network can also incorporate object parts in addition to the whole object. Better system performance is expected when considering image parts.

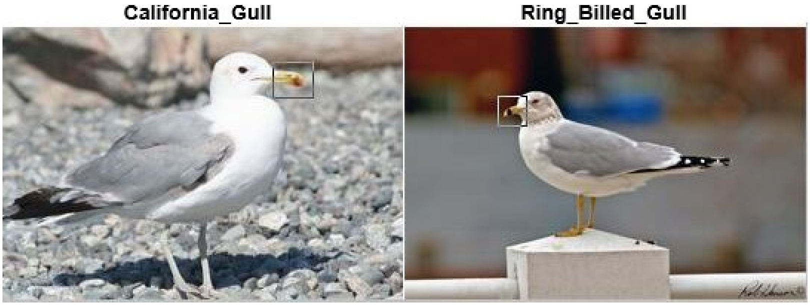

Fig.3. The salient difference between a California gull and a Ringed-billed gull lies in the pattern of their beaks.

Previous work [19] reported a bottom-up selective search method to generate part and object proposals, which used RCNN to perform object detection. In the experiments, they discovered that the region proposals are the bottleneck for precise fine-grained recognition. Salient differences among different fine-grained bird species are more likely to attach to some small parts. Once the crucial discriminative small parts are lost due to the unreliable proposal methods, it is hard for the sub-classification network to further distinguish them. For example, as shown in Fig.3, it is not straightforward to distinguish between a Ringed-billed gull and a California gull without identifying the pattern of their beaks. In our method, the Faster RCNN network can generate high quality proposals, since it exploits an effective proposal generation network RPN. RPN exploits a multi-task loss function used for classification and bounding-box regression of the translation-invariant anchors. The loss function is defined as:

$$L(\{p_i\},\{t_i\}) = \frac{1}{N_{cls}}\sum_i L_{cls}(p_i, p_i^*) + \lambda \frac{1}{N_{reg}}\sum_i p_i^* L_{reg}(t_i, t_i^*) \tag{1}$$

where i is the index of an anchor in a mini-batch and p_i is the predicted probability of anchor i being an object. The ground truth label $p_i = 1$ if the anchor is positive, and $p_i = 0$ if the anchor is negative. t_i is a vector representing the four parameterized coordinates of the predicted bounding box, and t_i^* is that of the ground truth box associated with a positive anchor. The classification loss L_{cls} is the log loss over the two classes (object vs. background). The regression loss function L_{reg} is of a robust L1 form, defined as:

$$L_{reg}(t_i, t_i^*) = \sum_i smooth_{L_1}(t_i, t_i^*) \tag{2}$$

$$smooth_{L_1}(x)\left\{\begin{matrix} 0.5x^2 & if\ |x|<1 \\ |x-0.5| & otherwise \end{matrix}\right\} \tag{3}$$

The two terms are normalized with N_{cls} and N_{reg}, and a balancing weight λ.

In our experiments, SFNet can achieve 82.0% accuracy only with average 10 high quality proposals per image, far less than thousands of bounding boxes produced from the selective search method [27].

3.2 Objective function

3.2.1. Softmax loss

Both Fast R-CNN and Faster RCNN drop the one-vs-rest SVM in the RCNN in order to obtain an end-to-end system. However, softmax loss encourages feature representation to learn inter-category variances instead of intra-category variances. This can be explained by the definition of softmax loss in Eqs. 4 and 5.

$$L = -\sum_{n=1}^{N}\sum_{c=1}^{C} t_{n,c}\log p_{n,c}, where\ p_{n,c} = \frac{e^{net_{n,c}}}{e^{\sum_{c=1}^{C} net_{n,c}}} \tag{4}$$

Denote $t_{n,c}$ *and* $p_{n,c}$ as the ground truth label and the predicted label for the nth sample and cth class. $t_{n,c} = 1$ if the nth sample belongs to the cth class, $t_{n,c} = 0$ otherwise. $net_{n,c}$ is the classification prediction from the neural network. Denote θ as the parameter of the network, the derivative is :

$$\frac{\delta L}{\delta \theta} = \sum_{n,c}(p_{n,c} - t_{n,c})\frac{\delta net_{n,c}}{\delta \theta} \tag{5}$$

Eq.6 shows that the number of the samples belonging to class c influences the gradient of the parameters. Suppose the prediction errors $p_{n,c} - t_{n,c}$ have similar magnitudes for all the samples, then we can infer that one class which has more samples, the magnitude of the gradient from it will be much larger than the magnitude of the gradient from the other classes. This results in the network parameters dominated by the class which has much more samples. Therefore, the existence of the dominated background samples (3/4 of all the training samples) leads to better feature representation for capturing inter-category variances.

3.2.2. One-vs-rest loss

For the Fast RCNN in the proposed framework, we replace the softamx loss function with a novel one-vs-rest loss, which is designed to capture variances among different subordinate categories. One-vs-rest loss consists of N (the number of subordinate categories) two-class cross entropy losses, and each two-class cross entropy loss function focuses on capturing the variances between one specific subordinate category and its similar categories. The objective function is the sum of N two-class cross entropy losses. At the training time, primitive detections with low scores, which are more likely to be the background, are discarded. This step is especially important since it makes one-vs-rest Fast RCNN network learn more discriminative features of different subordinate categories. Then each two-class cross entropy classifier is trained using the detections which have high scores on that specific category, as those high scored detections may be true positives or false positives (i.e. detections misclassified by SFNet whose ground truth labels are similar to that specific category). In this way, the negative training samples of each two-class cross entropy classifier are of the categories similar to the specific category, allowing each specific two-class cross entropy classifier to capture the variances between the specific category and its similar categories. At the test time, after non maximum suppression (NMS) operation on the primitive detections, less and higher quality detections are left. Then each of the left detections is again classified and regressed by the one-vs-rest Fast RCNN, and the output scores (N categories) are averaged (different from the multiply operation used in [30]) over the primitive scores in a category-by-category way to retrieve the final scores. Finally, the predicted label of the highest scored box is used as the predicted label of the whole image. The whole training process and the testing stage of RFNet are illustrated in Processes 1 and 2, accordingly.

Process 1: RFNet training process

Input: Ground truth labels and bounding boxes of the training set $GT = \{(L_1^*, B_1^*), \ldots, (L_N^*, B_N^*)\}$, B_i^* *and* $L_i^* \; (1 \le i \le N)$ denote the ground truth bounding boxes and its labels.

Output: Parameters of the SFNet w_{sf} and the one-vs-rest Fast RCNN w_{ovs} .

Step1: Fine-tune SFNet using GT and get the parameters of SFNet w_{sf} .

Step2: Pass the image x from training set through SFNet, and get M primitive detections $D = \phi_{w_{sf}}(x)$, ϕ is the SFNet function parameterized by w_{sf} . $D = \{(L_1, B_1, S_1), \ldots, (L_M, B_M, S_M)\}$, (L_i, B_i, S_i) are the predicted label, bounding box and score of the $i\,th \; (1 \le i \le M)$ primitive detection in image x .

Step3: Discard the primitive background detection (L_j, B_j, S_j) , if $S_j < \alpha, \alpha$ is a constant threshold.

Step4: Add primitive detection (L_j, B_j, S_j) into the training set of the $k\,th$ two-class cross entropy losses classifier O_{ovs}^k (responsible for classifying the $k\,th$ subordinate category), if $L_j = k$.

Step5: Train the kth two-class cross entropy losses classifier of one-vs-rest Fast RCNN network using the training samples in O_{ovs}^{k}, and obtain the final parameters of the one-vs-rest detection network w_{ovs}.

Process 2: RFNet testing process

Input: Image x in the testing set, parameters of the SFNet and the one-vs-rest Fast RCNN w_{sf} *and* w_{ovs}.

Output: label y of image x.

Step1: Pass image x through SFNet, and get N primitive detections $D=\phi_{w_{sf}}(x),\phi$ is well trained SFNet function parameterized by w_{sf} at the training stage above. $D=\{(B_1,S_1),\ldots,(B_N,S_N)\}$. B_j *and* S_j are the predicted bounding box and the score of the jth primitive detection in image x, here $S_j=(s_j^1,\ldots,s_j^k)$ is a $K-$dimensional vector, K is the number of classes($K=200$ in CUB-200-2011 dataset), each element s_j^k denotes the probability of the jth detection being an object of class $k, 1\le k\le K$.

Step2: Input image x and its N primitive detections D into the one-vs-rest Fast RCNN network. Get N refined detections and $D^{'}=\phi_{w_{ovs}}(x,D)$ corresponding to N primitive detections D. $D^{'}=\{(B_1^{'},S_1^{'}),\ldots,(B_N^{'},S_N^{'})\}, B_j^{'}$ *and* $S_j^{'}$ are the refined bounding box and the score of the jth primitive detection in image $x, S_j^{'}=(s_j^{'1},\ldots,s_j^{'K})$.

Step3: Computer the final score S_j^f of the jth detection as $S_j^f=((s_j^1+s_j^{'1})/2,\ldots,(s_j^K+s_j^{'K})/2), 1\le j\le N$. Update the score and the label of the jth detection $S_j^{'}=\max(s_j^f)$ *and* $L_j^{'}=\arg\max(S_j^f)$, then $P^{'}=\{(L_1^{'},B_1^{'},S_1^{'}),\ldots,(L_N^{'},B_N^{'},S_N^{'})\}$.

Step4: Finally, the image x ’ s label $y=L_i^{'}$ where $i=\arg\max(S_1^{'},\ldots,S_N^{'})$.

---

4 Experimental Results

4.1 Dataset

We evaluate the performance of our proposed framework for fine-grained recognition on CUB-200-2011 dataset [1], which is generally considered as the most extensive and competitive datasets in the literature. CUB-200-2011 contains 11,788 images of 200 bird species, each image has a single bounding box annotation, rough segmentations and 15 key points annotated, which is not used in our method.

4.2 Implementation details

The baseline models of our two networks are based on the VGG16 model [31], as done in current state-of-the-art methods [14, 15]. All the experiments are performed on a single NVIDIA K40 GPU. Parameters of the SFNet are initialized from the model pre-trained on the ImageNet dataset. Parameters of the one-vs-rest Fast RCNN are initialized from the SFNet model, and the new one-vs-rest loss layer is initialized from a Gaussian distribution.

4.3 Results and Comparisons

We first conduct some ablation experiments to analyse the cascaded structure and the one-vs-rest loss with regard to recognition performance, and then move on to the comparison against the previous work.

4.3.1. Ablation Experiments

Table 2. Recognition performance comparisons between SFNet, softmax RFNet and RFNet on CUB-200-2011, softmax RFNet consists of a standard Faster RCNN (SFNet) and a standard Fast RCNN with softmax loss.

Methods	Cascaded structure	One-vs-rest loss	Accuracy
SFNet			82.0%
Softmax RFNet	✓		82.9%
RFNet	✓	✓	84.0%

How important is the cascade structure? To evaluate the effectiveness of the cascaded structure, we compare SFNet with softmax RFNet, which consists of a standard Faster RCNN (SFNet) and a standard Fast RCNN with the softmax loss function. For softmax RFNet, the baseline model of the standard Fast RCNN is VGG16 and the parameters are initialized for the SFNet model as the same as RFNet. From Table 1, we observe that softmax RFNet improves accuracy by 0.9% over SFNet, and the experiment validates the effectiveness of the cascaded structure to eliminate the influence of excessive background samples during feature learning.

Sotfmax loss vs. One-vs-rest loss. The comparison between softmax RFNet and RFNet, shows that one-vs-rest loss improves accuracy by 1.1% over softmax loss. The results shown in Fig. 4 verify the ability of the one-vs-rest loss function of further capturing intra-category variances among the subordinate categories, and also reducing false positives mainly caused between ambiguous categories.

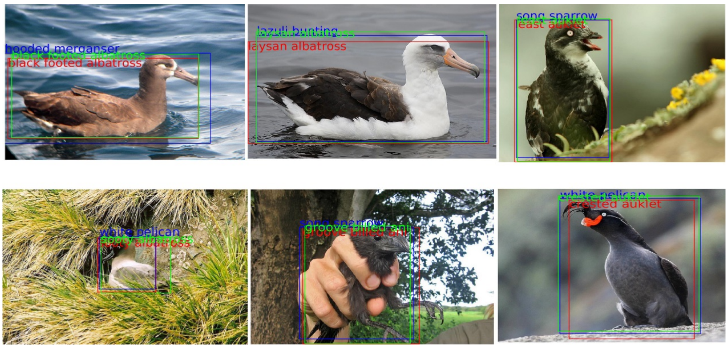

Fig. 4. Examples on the CUB-200-2011 dataset of SFNet detections (blue), RFNet detections (red) and ground truth bounding box (green). Images misclassified by SFNet are rectified by one-vs-rest Fast RCNN network.

4.3.2. Comparison with other state-of-the-art methods

This section shows the comparison results of our method against the previous work. For fair comparison, we report the results with varying degrees of supervision such as part annotation or bounding-boxes at the training and the testing time.

Table 3. Recognition performance comparisons of the current state of the art methods on CUB-200-2011, sorted by the amount of annotation used. RFNet refers to our unified cascade detection framework. “Parts” refers to using any annotation at the level of parts at all. “BBox” and “Parts” refer to any annotation at the level of bounding box and part separately.

Method	Train Anno.	Test Anno.	Acc.
Alignment[32]	n/a	n/a	53.6%
Attention[24]	n/a	n/a	77.9%
NAC[25]	n/a	n/a	81.0%
Bilinear[33]	n/a	n/a	84.1%
No parts[20]	BBox	n/a	82.0%
Our RFNet	BBox	n/a	84.0%
Alignment[32]	BBox	BBox	67.0%
No parts[20]	BBox	BBox	82.8%
PS-CNN[21]	BBox+Parts	BBox	76.6%
Deep LAC[18]	BBox+Parts	BBox	80.2%
SPDA[15]	BBox+Parts	BBox	84.55%
FOAF[34]	BBox+Parts	BBox+Parts	81.2%
Part RCNN[19]	BBox+Parts	BBox+Parts	82.0%
PN-CNN[14]	BBox+Parts	BBox+Parts	85.4%

The comparison results illustrated in Table 2 show that our RFNet performs much better than the previous unsupervised methods [24, 25, 32], and outperforms part-based methods [18, 19, 21, 34]. RFNet also achieves comparable performance against the state-of-the-art, part-free, fine-grained recognition method [33]. [33] presents bilinear models that exploit two CNNs to extract features while we use a single cascaded structure to extract features which is easier to train. However, our method is slightly worse than the current state-of-the art methods [14, 15], due to the significant advantage of exploring part information for bird recognition. [32] is with box level annotation at both the training and testing stages, and achieves about 13.4% higher accuracy than that without any annotation. [20] introduced box level annotation at the testing time, and also achieved better performance. All these developments verify that leveraging more additional supervision results in higher performance. It is worth emphasizing that RFNet improves the detection and the loss layers for better feature learning. We anticipate that leveraging part annotations in our cascade detection framework will result in higher performance due to the additional supervision.

5 Conclusion and Discussion

In this paper, we have proposed a novel cascade detection framework for fine-grained recognition tasks without considering parts. The proposed cascaded detection framework is well adapted for fine-grained recognition by introducing a one-vs-rest loss function, which can capture more intra-category variances. Experiments showed that our proposed recognition framework achieved comparable performance against the other state-of-the-art part free fine-grained recognition methods on the CUB-200-2011 Birds dataset.

The cascaded framework boosts the classification accuracy, but the two networks are trained respectively and cannot meet the requirement of many real-time applications. Taking into account the speed of the proposed framework, and introducing the proposed solution to applications such as surveillance systems and the recommendation of relevant products in e-commerce become one of the future research directions.

6 Acknowledgment

2014DFA10410. H. Zhou is supported by UK EPSRC under Grant EP/N011074/1 and Royal Society-Newton Advanced Fellowship under Grant NA160342.

7 References

[1] Wah, C., Branson, S., Welinder, P., Perona, P., & Belongie, S. (2011). The caltech-ucsd birds-200-2011 dataset.

[2] Branson, S., Van Horn, G., Wah, C., Perona, P., & Belongie, S. (2014). The ignorant led by the blind: A hybrid human–machine vision system for fine-grained categorization. International Journal of Computer Vision, 108(1-2), 3-29.

[3] Berg, T., Liu, J., Lee, S. W., Alexander, M. L., Jacobs, D. W., & Belhumeur, P. N. (2014, June). Birdsnap: Large-scale fine-grained visual categorization of birds. In Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on (pp. 2019-2026). IEEE.

[4] Sfar, A. R., Boujemaa, N., & Geman, D. (2013, June). Vantage feature frames for fine-grained categorization. In Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on (pp. 835-842). IEEE.

[5] Kumar, N., Belhumeur, P. N., Biswas, A., Jacobs, D. W., Kress, W. J., Lopez, I. C., & Soares, J. V. (2012). Leafsnap: A computer vision system for automatic plant species identification. In Computer vision–ECCV 2012 (pp. 502-516). Springer, Berlin, Heidelberg.

[6] Maji, S. (2012, October). Discovering a lexicon of parts and attributes. In European Conference on Computer Vision (pp. 21-30). Springer, Berlin, Heidelberg.

[7] Li, Z., Nie, F., Chang, X., & Yang, Y. (2017). Beyond Trace Ratio: Weighted Harmonic Mean of Trace Ratios for Multiclass Discriminant Analysis. IEEE Transactions on Knowledge and Data Engineering, 29(10), 2100-2110.

[8] Yu, C., J. Li, X. Li, X. Ren, and B. Gupta, Four-image encryption scheme based on quaternion Fresnel transform, chaos and computer generated hologram. Multimedia Tools and Applications, 2017: p. 1-24.

[9] Chang, X., & Yang, Y. (2017). Semi-supervised feature analysis by mining correlations among multiple tasks. IEEE transactions on neural networks and learning systems, 28(10), 2294-2305.

[10] Zhang, Z., R. Sun, C. Zhao, J. Wang, C.K. Chang, and B.B. Gupta, CyVOD: a novel trinity multimedia social network scheme. Multimedia Tools and Applications, 2017. 76(18): p. 18513-18529.

[11] Ibtihal, M. and N. Hassan, Homomorphic Encryption as a Service for Outsourced Images in Mobile Cloud Computing Environment. International Journal of Cloud Applications and Computing (IJCAC), 2017. 7(2): p. 27-40.

[12] Atawneh, S., A. Almomani, H. Al Bazar, P. Sumari, and B. Gupta, Secure and imperceptible digital image steganographic algorithm based on diamond encoding in DWT domain. Multimedia tools and applications, 2017. 76(18): p. 18451-18472.

[13] Jouini, M. and L.B.A. Rabai, A Security Framework for Secure Cloud Computing Environments. International Journal of Cloud Applications and Computing (IJCAC), 2016. 6(3): p. 32-44.

[14] Branson, S., G. Van Horn, S. Belongie, and P. Perona, Bird species categorization using pose normalized deep convolutional nets. arXiv preprint arXiv:1406.2952, 2014.

[15] Zhang, H., Xu, T., Elhoseiny, M., Huang, X., Zhang, S., Elgammal, A., & Metaxas, D. (2016). Spda-cnn: Unifying semantic part detection and abstraction for fine-grained recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 1143-1152).

[16] Chang, X., Ma, Z., Lin, M., Yang, Y., & Hauptmann, A. G. (2017). Feature interaction augmented sparse learning for fast kinect motion detection. IEEE Transactions on Image Processing, 26(8), 3911-3920.

[17] Chang, X., Z. Ma, Y. Yang, Z. Zeng, and A.G. Hauptmann, Bi-level semantic representation analysis for multimedia event detection. IEEE transactions on cybernetics, 2017. 47(5): p. 1180-1197.

[18] Lin, D., Shen, X., Lu, C., & Jia, J. (2015, June). Deep lac: Deep localization, alignment and classification for fine-grained recognition. In Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on (pp. 1666-1674). IEEE.

[19] Zhang, N., Donahue, J., Girshick, R., & Darrell, T. (2014, September). Part-based R-CNNs for fine-grained category detection. In European conference on computer vision (pp. 834-849). Springer, Cham.

[20] Krause, J., Jin, H., Yang, J., & Fei-Fei, L. (2015, June). Fine-grained recognition without part annotations. In Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on (pp. 5546-5555). IEEE.

[21] Huang, S., Xu, Z., Tao, D., & Zhang, Y. (2016). Part-stacked cnn for fine-grained visual categorization. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 1173-1182).

[22] Chang, X., Yu, Y. L., Yang, Y., & Xing, E. P. (2017). Semantic pooling for complex event analysis in untrimmed videos. IEEE transactions on pattern analysis and machine intelligence, 39(8), 1617-1632.

[23] Alsmirat, M.A., Y. Jararweh, M. Al-Ayyoub, M.A. Shehab, and B.B. Gupta. Accelerating compute intensive medical imaging segmentation algorithms using hybrid CPU-GPU implementations. Multimedia Tools and Applications, 2017. 76(3): p. 3537-3555.

[24] Xiao, T., Xu, Y., Yang, K., Zhang, J., Peng, Y., & Zhang, Z. (2015, June). The application of two-level attention models in deep convolutional neural network for fine-grained image classification. In Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on (pp. 842-850). IEEE.

[25] Simon, M., & Rodner, E. (2015). Neural activation constellations: Unsupervised part model discovery with convolutional networks. In Proceedings of the IEEE International Conference on Computer Vision (pp. 1143-1151).

[26] Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 580-587).

[27] Uijlings, J.R., K.E. Van De Sande, T. Gevers, and A.W. Smeulders, Selective search for object recognition. International journal of computer vision, 2013. 104(2): p. 154-171.

[28] Girshick, R. (2015). Fast r-cnn. arXiv preprint arXiv:1504.08083.

[29] Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. In Advances in neural information processing systems (pp. 91-99).

[30] Yang, B., Yan, J., Lei, Z., & Li, S. Z. (2016). Craft objects from images. arXiv preprint arXiv:1604.03239.

[31] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556. 2014 Sep 4.

[32] Gavves, E., Fernando, B., Snoek, C. G., Smeulders, A. W., & Tuytelaars, T. (2013, December). Fine-grained categorization by alignments. In Computer Vision (ICCV), 2013 IEEE International Conference on (pp. 1713-1720). IEEE.

[33] Lin, T. Y., RoyChowdhury, A., & Maji, S. (2015). Bilinear cnn models for fine-grained visual recognition. In Proceedings of the IEEE International Conference on Computer Vision (pp. 1449-1457).

[34] Zhang, X., Xiong, H., Zhou, W., & Tian, Q. (2014, November). Fused one-vs-all mid-level features for fine-grained visual categorization. In Proceedings of the 22nd ACM international conference on Multimedia (pp. 287-296). ACM.